

Bioinformatika III

Trimačių struktūrų analizė ir spėjimas

Paskaita 2 – struktūrinių failų formatai (CIF)

Saulius Gražulis
2011 m.

CIF ir mmCIF formatas

- ASCII koduotės tekstiniai failai
- Laisvo formato sintaksė
- Duomenys žymimi raktiniais žodžiais, bet įrašai nėra suskirstyti eilutėmis
- Duomenų raktiniai žodžiai ir jų semantika aprašyti CIF žodynuose (CIF dictionaries)

<http://www.iucr.org/iucr-top/cif/standard/cifstd1.html>

<http://www.iucr.org/iucr-top/cif/spec/version1.1/cifsyntax.html>

<http://www.csse.uwa.edu.au/star/cif-bnf.html>

<http://ugo.crl.nitech.ac.jp/~ddb/CIF/skunkworks/html/index.html>

<http://ugo.crl.nitech.ac.jp/~ddb/CIF/skunkworks/html/star/spec/index.html>

<http://ugo.crl.nitech.ac.jp/~ddb/CIF/skunkworks/html/cif/spec/index.html>

CIF failo pavyzdys

```
data_1KNV
#
_entry.id      1KNV
#
_audit_conform.dict_name      mmcif_pdbx.dic
_audit_conform.dict_version   1.044
...
_cell.entry_id      1KNV
_cell.length_a      121.230
_cell.length_b      122.280
_cell.length_c      56.870
_cell.angle_alpha    90.00
_cell.angle_beta     90.00
_cell.angle_gamma    90.00
...
loop_
_atom_site.group_PDB
_atom_site.id
_atom_site.type_symbol
...
ATOM      1      N  N      . ASN A 1 4      ? 3.407  40.303 50.109  1.00 66.19 ? ? ? ? ? 4      ASN A N      1
ATOM      2      C  CA      .
ASN A 1 4      ? 4.752
              40.029 49.523  1.00 67.25 ? ? ? ? ? 4      ASN A CA      1
```

Kalbos

Matematikoje, *kalba* L vadinama pora $(A, W \subset A^*)$, kur:

A yra *baigtinis* alfabetas (t.y. baigtinis simbolių rinkinys),

A^* yra (begalinė) visų galimų *baigtinių* simbolių eilučių iš A aibė,

W yra A^* poaibis.

Gramatikos

Gramatikos pavyzdys:

$$R \rightarrow S \mid S + R \mid S - R$$

$$S \rightarrow D \mid D \times S \mid D / S$$

$$D \rightarrow V \mid (R)$$

$$V \rightarrow a \mid b \mid c$$

Teisingas sakiny:

$$(a + b) / (a - b \times b / c)$$

Neteisingas sakiny:

$$((a)(+ - b) / (() a - b b b c d e f \times b / c)$$

Bekaus-Nauro forma (angl. Backus-Naur Form (BNF))

```
<reiškinys> ::= <sandauga>  
              | <sandauga> + <reiškinys>  
              | <sandauga> - <reiškinys>
```

```
<sandauga> ::= <daugiklis>  
              | <daugiklis> * <sandauga>  
              | <daugiklis> / <sandauga>
```

```
<daugiklis> ::= <vardas> | ( <reiškinys> )
```

```
<vardas> ::= a | b | c
```

CIF failo sintaksè, STAR formatas

Gramatika Bekuso-Nauro (Backus-Naur) forma:

```
...
<data_block> ::= <data_heading> <data>+ { <wspace>+ | <EOF> }
<data_heading> ::= <DATA_> <non_blank_char>+
<data> ::= { <wspace>+ <data_name> <wspace>* <blank>
             <data_value_1> }
           | { <wspace>+ <data_name> <wspace>* <terminate>
             <data_value_2> }
           | <data_loop>

<data_loop> ::= <wspace>+ <LOOP_> <data_loop_field> <data_loop_values>

<data_loop_field> ::= { <wspace>+ <data_name> }+
<data_name> ::= ' ' <non_blank_char>+
<data_loop_values> ::= { { <wspace>* <blank> <data_value_1> }
                       | { <wspace>* <terminate> <data_value_2> } }+
...

```

CIF failo sintaksės ypatumai

```
# Komentarai prasideda "groteliu" (#) simboliu
# Leistini tik ASCII simboliai, todėl tenka rasyti "sveplai"

data_DataName
_tag1 value
_tag2 1.23(3)
_tag3 'eilutes su tarpais turi būti viengubose kabutėse'
_tag4 "arba dvigubose kabutėse"
_tag5 'žodis d'Alamber (su kabute-apostrofu) gali būti kabučių vidury (!)'
_tag5a
'reikšmė gali būti bet kur, net ir kitoje eilutėje'

loop_
_tag6 # duomenų žymių išdėstymas eilutes bet koks
_tag7 _tag8
123 456 789
111 222
333

DaTa_NextDataName # pagal gramatiką, data_ neskiria didžiųjų ir mažųjų
_tag1 123 # duomenų žymės unikalios data_ bloke, bet skirtinguose
          # blokuose gali kartotis

# pabaigoje jokios žymės
```


CIF failo semantika, CIF žodynai

Ką reiškia '_atom_site_label'? Koks žymuo (tag) naudojamas koordinatėms?

```
data_atom_site_fract_
  loop_ _name
    '_atom_site_fract_x'
    '_atom_site_fract_y'
    '_atom_site_fract_z'
  _category
  _type
  numb
...
  _list_reference
  _definition
;           Atom-site coordinates as fractions of the _cell_length_ values.
;

data_atom_site_label
  _name
  _category
  _type
  char
...
_definition
;           The _atom_site_label is a unique identifier for a particular site
           in the crystal.
...
```

CIF “žodynų žodynai”, DDL

Ką reiškia '_name'? Koks žymuo nurodo duomenų tipą?

```
data_name
  _definition
;      The data name(s) of the defined item(s). If data items are
      closely related or represent an irreducible set, their names
      may be declared as a looped sequence in the same definition.
;
  _name          '_name'
  _category      name
  _type          char
  _list          both
  loop_ _example '_atom_site_label'
                '_atom_attach_all'  '_atom_attach_ring'
```

mmCIF, DDL2 žodynas

- Standartinis CIF žodynas neturėjo mechanizmų makromolekulių struktūroms užrašyti
- PDB sukūrė naują mmCIF (macromolecular CIF) žodyną, pritaikytą biologinėms makromolekulėms

(mm)CIF formato **privalumai**

- ASCII tekstas, įskaitomas tiek žmogui, tiek mašinai
- Griežtai apibrėžta sintaksė
- Semantinė informacija kaupiama to paties formato žodynuose
- Tinka visiems struktūrinės informacijos tipams
- Numatyta standartinė galimybė įvesti naujus duomenų laukus

(mm)CIF formato trūkumai

- Sudėtinga gramatika
- Daugelį dažnai pasitaikančių klaidų sunku griežtai lokalizuoti ar net aptikti, ypač lentelėse ('loop_' loops)
- Skaitymas reikalauja sudėtingo specializuoto sintaksinio analizatoriaus
- Kai kurie semantiniai aspektai iki šiol nepakankamai griežtai aprašyti
- Silpnas daugianacionalinių simbolių palaikymas, nepalaikomas UTF-8
- Pradinis CIF formatas netinka makromolekulėms, reikalingas dar sudėtingesnis mmCIF formato variantas (žodynas)
- Dėl didelio variantų skaičiaus skaitančios programos yra sudėtingos, ir ne visi *teisingi* CIF failai bus skaitomi visų programų

PDB XML schema

- mmCIF žodynus galima paversti 1:1 į XML schema

```
<?xml version="1.0" encoding="UTF-8" ?>  
<PDBx:datablock datablockName="1KNV"  
  xmlns:PDBx="http://deposit.pdb.org/pdbML/pdbx.xsd"  
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"  
  xsi:schemaLocation="http://deposit.pdb.org/pdbML/pdbx.xsd pdbx.xsd">
```

...

```
<PDBx:atom_siteCategory>  
  <PDBx:atom_site id="1">  
    <PDBx:group_PDB>ATOM</PDBx:group_PDB>  
    <PDBx:type_symbol>N</PDBx:type_symbol>  
    <PDBx:label_atom_id>N</PDBx:label_atom_id>  
    <PDBx:label_alt_id xsi:nil="true" />  
    <PDBx:label_comp_id>ASN</PDBx:label_comp_id>  
    <PDBx:label_asym_id>A</PDBx:label_asym_id>  
    <PDBx:label_entity_id>1</PDBx:label_entity_id>  
    <PDBx:label_seq_id>4</PDBx:label_seq_id>  
    <PDBx:Cartn_x>3.407</PDBx:Cartn_x>  
    <PDBx:Cartn_y>40.303</PDBx:Cartn_y>  
    <PDBx:Cartn_z>50.109</PDBx:Cartn_z>
```

...

Ontologijos ir semantiniai tinklai

- Ontologija (Graikiškai ων „būtis“, λόγος „žodis“ ar „kalba“) — filosofijos skyrius, /.../ Pagrindinis ontologijos klausimas — „Kas egzistuoja?“
- Ontologija — kompiuterijoje šiuo termino daugiskaitine forma „ontologijos“ ... vadinamas tam tikros srities sąvokų visumos specifikavimas išreikštu pavidalu.

<http://lt.wikipedia.org/wiki/Ontologija>

[http://lt.wikipedia.org/wiki/Ontologija_\(informatika\)](http://lt.wikipedia.org/wiki/Ontologija_(informatika))

Kaip turėtų atrodyti „idealus“ formatas?

- Tekstinis, ASCII -> UTF8
- Įrašas <=> eilutė
- Laukai atskirti tarpais
- Raktiniai žodžiai nurodo įrašą
- Fiksuoti įrašų laukai ir tipai?
- Jokių dydžio apribojimų!

„Idealaus“ formato pavyzdys...

```
FORMAT My ideal macromolecular data format ver. 0.0
```

```
#
```

```
# Komentarai gali būti skirti žmogui
```

```
#
```

```
TITLE Restrikcijos endonukleazės struktūra
```

```
AUTHORS Saulius Gražulis; Elena Manakova (Манакова, Елена)
```

```
CELL 100.0 100.0 100.0 100.0 90 90 90
```

```
SPACEGROUP P212121
```

```
#
```

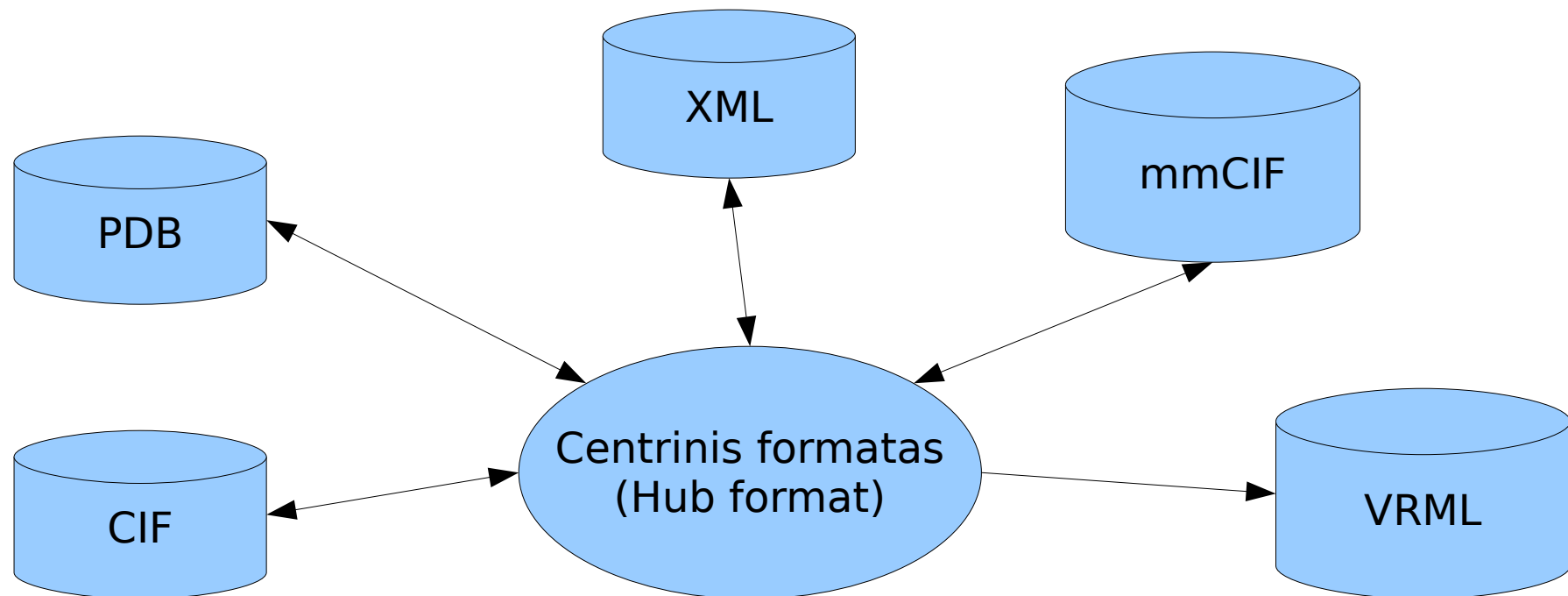
```
ATOM N ASN A 4 3.407(1) 40.303(2) 50.109(11) 1.00 66.19 N
```

```
ATOM CA ASN A 4 4.752 40.029 49.523 1.00 67.25 C
```

```
...
```

Galimi „nuosavo“ formato panaudojimai

- „Centriniai“ formatai (Hub formats)



Sėkmingo Hub-formato pavyzdys: netpbm <http://netpbm.sourceforge.net/>